Sentiment Analysis of Arabic Tweets: Opinion Target Extraction

Salima BEHDENNA, Fatiha Barigou, Ghalem Belalem
Computer Science Department, Faculty of Sciences, University of Oran 1 Ahmed Ben Bella
Oran, Algeria
behdennasalima@gmail.com
fatbarigou@gmail.com
ghalem1dz@gmail.com

ABSTRACT: Due to the increased volume of Arabic opinionated posts on different social media, Arabic sentiment analysis is viewed as an important research field. Identifying the target or the topic on which opinion has been expressed is the aim of this work. Opinion target identification is a problem that was generally very little treated in Arabic text. In this paper, an opinion target extraction method from Arabic tweets is proposed. First, as a preprocessing phase, several feature forms from tweets are extracted to be examined. The aim of these forms is to evaluate their impacts on accuracy. Then, two classifiers, SVM and Naïve Bayes are trained. The experiment results show that, with 500 tweets collected and manually tagged, SVM gives the highest precision and recall (86%).

Subject Categories and Descriptors

H.3.1 [Content Analysis and Indexing] H.3.3 [Information Search and Retrieval]

General Terms: Data Mining, Sentiment Analysis, Twitter, Arabic Opinion Processing, Arabic Text Mining

Keywords: Opinion Mining, Arabic Sentiment Analysis, Opinion Target, Machine Learning, Arabic Tweet

Received: 12 April 2018, Revised 3 June 2018, Accepted 12 June 2018

DOI: 10.6025/jdim/2018/16/6/324-331

1. Introduction

Due to the emergence of Web2.0, users can share their opinions and sentiments on a variety of topics in new interactive forms where users are not only passive information receivers.

Sentiment analysis or opinion mining is the computational study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes (Liu and Zhang, 2012). The aim of sentiment analysis is to automatically extract users' opinions. The main tasks of sentiment analysis (SA) are (Elarnaoty, 2012):

- Subjectivity extraction.
- Opinion polarity identification.
- Opinion element's extraction task.
- Development of resources like sentiment lexicon and annotated corpora required for previous tasks.

With the increase in the volume of Arabic opinionated posts on different social media, Arabic Sentiment Analysis (ASA) is viewed as an important research field. Most of researches in Arabic sentiment Analysis attempt to determine the overall opinion polarity. This work focuses on opinion target extraction subtask, which is a subject that has been little studied to date for ASA. This task

aims to extract topics expressed within opinions. For example, when saying in an Arabic opinion; "سوني 4 أفضل جهاز" ("Sony 4 the best device"), the opinion target is "سوني 4" ("Sony 4").

In this paper, we propose a method for extracting opinion targets from Arabic tweets by modeling the problem of Opinion Target Extraction as a machine learning classification task and combining a number of the available resources for Arabic language together with tweets features.

The rest of the paper is organized as follows. Section 2 discusses related work, section 3 describes the features and the method for extracting the opinion target. Section 4 focuses on the experimental results and discussion. Section 5 concludes the paper.

2. Related Work

Many works have focused on the task of opinion target extraction from documents in English. For example, in (Hu and Liu, 2004), the authors explored the problem of generating feature-based summaries of customer reviews of products sold online and proposed a lexicon-based algorithm. Features are extracted using association rule mining and POS tags. In the work of (Ding et al., 2009), the authors studied the entity discovery, and the entity assignment problems. They applied automatic pattern extraction based on POS tags and a starting seed patterns, then assigning entities based on pattern matching. (Li et al., 2012) modeled aspect extraction as a shallow semantic parsing problem. A parse tree is built for each sentence and structured syntactic information within the tree is used to identify aspects. (Shang et al. 2012) proposed a new method to extract opinion targets for short comments by developing a two-dimensional vector representation for words and a back propagation neural network for classification. (Liu et al., 2015) applied a word alignment model in order to extract opinion targets. Then, a graph-based co-ranking algorithm is exploited to estimate the confidence of each candidate. Finally, candidates with higher confidence are extracted as opinion targets or opinion words.

Recent methods using deep learning approaches show performance improvement on standard datasets. (Poria et al., 2016) combined convolutional neural network and

linguistic patterns for aspect extraction. (Yin et al., 2016) developed a novel approach to aspect term extraction based on unsupervised learning of distributed representations of words and dependency paths. This method leverages the dependency path information to connect words in the embedding space. The embeddings can significantly improve the CRF based aspect term extraction. (Wang et al., 2017) proposed a multi-layer attention network CMLA, for aspect-opinion co-extraction which does not require any parsers or linguistic resources.

The works for opinion target extraction for Arabic has not been widely investigated. In (Alkadri and ElKorany, 2016), the authors propose a feature-based opinion mining framework for Arabic reviews. This framework uses the semantic of ontology and lexi-cons in the identification of opinion target and their polarity. In (Ismail et al., 2016) proposed a generic approach that extracts the entity aspects and their attitudes for reviews written in modern standard Arabic. A two-stage method for annotating targets of opinions was developed in (Farra et al., 2015) by using the crowdsourcing tool Amazon Mechanical Turk. The first stage consists of identifying candidate targets "entities" in a given text. The second stage consists of identifying the opinion polarity (positive, negative, or neutral) expressed about a specific entity.

3. Proposed Method

This work focuses on the opinion target extraction as part of the sentiment analysis task. We model the problem as a machine learning classification task. The process which follows the data mining process is composed of four steps:

3.1. Corpus Building

We need an annotated Arabic corpus for opinion target. Unfortunately, Arabic corpus required for this task is not available, therefore we decide to build our own Arabic corpus, and manually annotate it for opinion target.

We used the **Twitter Archive Google Spreadsheet** (TAGS) to collect tweets related to opinions expressed in Arabic from the topic: "*mobile phone brand*". After filtering out retweets and performing some pre-processing steps to clean up unwanted content like URL, we ended up with 500 tweets. We then manually annotated the opinion target in these tweets. Table 1 shows some examples of manually annotated tweets.

	Tweet	Target
1	بطارية # أبل ذات جودة خيالية	بطارية
2	افضل جهاز # سوني	سوني
3	السماعة # في عيب وجود عن يصرح سامسونج	سماعة
4	شحن أسرع مع شاحن # أبل	شاحن

Table 1. Examples of annotated Tweets

3.2. Preprocessing

In this phase, every tweet is tokenized into words and several forms are generated to be examined. The aim of these forms is to evaluate their impacts on accuracy. The following figure (Figure 1.) shows several forms of pre-processed tweets.

- Form (a): After removing the stop words and special characters, every tweet is transformed into a feature vector which includes all the remaining words.
- Form (b): After removing the special characters like (RT, URL, @, #), words are stemmed using the Khoja stemmer (Khoja and Garside, 1999) and combined with

stop words removal before transforming into a feature vector.

- Form (c): consists of form (b) followed by filtering of words according to their grammatical categories. We have retained the nouns and adjectives. The words are tagged using Stanford Arabic part of speech tagger².

At the end of this preprocessing step, tweets data are converted from text format into ARFF format required by the WEKA tool³. The tool we have used for classification step.

Figure 2 shows an example of tweets and the different forms of preprocessing.

Figure 1. Tweets Preprocessing

Figure 2. Example of different forms of tweets

¹ <https://tags.hawksey.info/>

² <http://nlp.stanford.edu/software/tagger.shtml>

³ <http://www.cs.waikato.ac.nz/ml/weka>

Number of tweets	100	200	300	400	500
Words (form a)	455	714	1066	1214	1256
Stem (form b)	426	654	942	1065	1097
Stem and POS (form c)	317	494	719	815	839

Table 2. Size of the feature Vector

The size of the feature vector depends on the number of tweets and the different forms used in preprocessing as shown in Table 2.

3.3. Classification

In the literature, several machine learning techniques are used, but two of them appear to provide the best results. They are the SVM and NB classifiers (Behdenna et al., 2016). The data mining tool we have used is Weka 3.7 open-source data mining software.

4. Experiments and Discussions

During all the experiments that we have carried, we used 10-fold cross-validation to train the classifiers (SVM and

NB). Each dataset is divided into 10 parts; one is used for testing and 9 for training in the first run. This process is repeated 10 times, using a different testing fold in each case. Experiments are carried out with different forms of preprocessed tweets.

The performance of classification model is measured by evaluating the precision, recall and F-measure. They are measured with equations (1), (2) and (3).

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

	Precision		Recall		F-measure	
	NB	SVM	NB	SVM	NB	SVM
100 tweets	0.550	0.534	0.630	0.640	0.573	0.545
200 tweets	0.557	0.541	0.650	0.675	0.590	0.572
300 tweets	0.648	0.663	0.680	0.705	0.641	0.661
400 tweets	0.658	0.673	0.697	0.720	0.663	0.682
500 tweets	0.672	0.680	0.701	0.741	0.676	0.691

Table 3. Classifier performance using simple words and according to the size of the dataset

Figure 3. Experiment 1 with NB

$$F\text{-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

4.1 Experiment 1: Impact of Simple Words by Varying the Corpus Size

The first experiment is carried out to evaluate the effect of using simple words on the performance of SVM and NB classifiers. Table 3 shows the performance obtained from each classifier.

As shown in figure 3 and figure 4, the passing from 100 tweets to 500 tweets improve significantly the performance. In the case of NB: Precision increased 12%, Recall increased 7.1%, and F-measure increased 10,3%. In the case of SVM: Precision increased 14%, Recall increased 10%, and F-measure increased 14,6%.

4.2. Experiment 2: Impact of stemming

This experiment is carried out to evaluate the effectiveness of stemming in the classification process.

Figure 4. Experiment 1 with SVM

	Precision		Recall		F-measure	
	NB	SVM	NB	SVM	NB	SVM
100 tweets	0653	0552	066	064	0631	0568
200 tweets	0742	0756	072	076	0707	0736
300 tweets	0727	0756	0727	0777	0711	076
400 tweets	0775	081	0773	0813	0768	0805
500 tweets	0821	086	0814	086	0816	0857

Table 4. Classifier performance using stemming by varying dataset size

Figure 5. Experiment 2 with NB

Figure 6. Experiment 2 with SVM

As illustrated in table 4, figure 5 and figure 6:

- Both the NB and SVM classifiers performed best compared to experiment 1. This means that stemming enables to generate a representative feature vector of tweets. For SVM, we observe a performance improvement in F-measure from 69, 1% (using simple words) to 85.67% for 500 tweets. And an improvement in F-measure from 67, 6% to 81, 6% for NB.

- The performance results of the two classifiers are close but SVM provides the best results.

4.3 Experiment 3: Impact of Stemming and Part of Speech Tagging

This experiment was carried out to assess the effect of stemming followed by filtering of words according to their grammatical categories. We have retained only nouns and adjectives, we have noticed that noun phrases are regarded as opinion target candidates.

	Precision		Recall		F-measure	
	NB	SVM	NB	SVM	NB	SVM
100 tweets	0,659	0,605	0,66	0,63	0,638	0,603
200 tweets	0,659	0,713	0,66	0,735	0,637	0,702
300 tweets	0,655	0,662	0,66	0,693	0,648	0,662
400 tweets	0,719	0,673	0,725	0,68	0,709	0,67
500 tweets	0,803	0,76	0,8	0,708	0,797	0,701

Table 5. Classifier performance using stemming and POS by varying dataset size

Figure 7. Experiment 3 with NB

Figure 8. Experiment 3 with SVM

As illustrated in table 5, figure 7 and figure 8:

- Unlike previous experiments, we observe a performance drop in precision from 86% to 70.6% for SVM.
- The best results are acquired using 500 tweets.

- The classifier NB perform better than SVM.

4.4 Discussions

In the three experiments, results show that the best results are acquired when the corpus of 500 tweets is used. Table 6 shows the precision, recall and F-measure obtained from each classifier for 500 tweets.

	Precision		Recall		F-measure	
	NB	SVM	NB	SVM	NB	SVM
Simple word	0,672	0,68	0,701	0,741	0,676	0,691
Stemming	0,821	0,86	0,814	0,86	0,816	0,857
Stemming & POS	0,803	0,706	0,8	0,708	0,797	0,701

Table 6. Classifier comparison using 500 tweets

Figure 9. Comparaison

As illustrated in table 6 and figure 9, we can state the following findings:

- Both the NB and SVM classifiers performed best when trained on sufficiently large corpus.
- The SVM performs the best when using simple words or stemming. When considering a corpus of 500 tweets, SVM slightly improves the performance when using simple words of 1% in terms of precision, 4% in terms of recall, 1.5% in terms of F-measure, and when using stemming SVM exceeds NB of 4.9% in terms of precision, 4.6% in terms of recall, and 4.1% in terms of F-measure
- Stemming combined with POS adversely affected the performance of classification.
- NB perform better than SVM in the case of stemming combined with the grammatical filtering of 9.7% in terms of precision, 9.2% in terms of recall, and 9.6% in terms of F-measure.
- By comparing the performance results of both classifiers, we see that using stem words as features give better classifier performance compared with other types of features (simple words and POS).

5. Conclusion

In this paper, we propose a method to extract opinion target from Arabic tweets. For this goal, we employed SVM and NB classifiers. The feature vectors of the tweets are preprocessed in several ways and the effects of these features on the classifiers' accuracy were investigated. The comparison is based on standard accuracy measures such as precision, recall, and F-measure. The results showed that, with 500 tweets collected and manually tagged, stemming combined with stop words removal improved the performance of classification.

We cast the problem of Opinion Target Extraction as a machine learning classification task. Hence to obtain the best results, this method requires a large corpus to allow better learning.

In the future work, a larger corpus will be used and we intend to use deep learning approaches.

References

- [1] Alkadri, A. M., ElKorany, A. M. (2016). Semantic Feature Based Arabic Opinion Mining Using Ontology. *International Journal of Advanced Computer Science and Applications*, 7(5) 577-583.
- [2] Behdenna, S., Barigou, F., Belalem, G. (2016). Sentiment Analysis at Document Level. *In: Proc of International Conference on Smart Trends for Information Technology and Computer Communications*, Singapore

p. 159-168. Springer.

- [3] Ding, X., Liu, B., Zhang, L. (2009). Entity discovery and assignment for opinion mining applications. *In: Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 125-1134, ACM, August, 2016.
- [4] Elarnaoty, M. (2012). A Machine Learning Approach for Opinion Holder Extraction in Arabic Language, *International Journal of Artificial Intelligence & Applications*, 3 (2) 45–63.
- [5] Farra, N., McKeown, K. & Habash, N. (2015). Annotating Targets of Opinions in Arabic using Crowdsourcing. *In: ANLP Workshop 2015* (p. 89). (July).
- [6] Hu, M., Liu, B. (2004). Mining and summarizing customer reviews. *In: Proc. of the Tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 168–177, ACM, August 2004.
- [7] Ismail, S., Alsammak, A., Elshishtawy, T. (2016). A Generic Approach for Extracting Aspects and Opinions of Arabic Reviews. *In: Proc of the 10th International Conference on Informatics and Systems* p. 173-179. ACM. 2016.
- [8] Khoja, S., Garside, R. (1999). Stemming Arabic Text, Computing Department, Lancaster University, Lancaster, UK. <http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps>
- [9] Liu, B., Zhang, L. (2012). A survey of opinion mining and sentiment analysis. *In: Mining text data*. p. 415-463. Springer US.
- [10] Li, S., Wang, R., Zhou, G. (2012). Opinion target extraction using a shallow semantic parsing framework. *In: ProcTwenty-sixth AAAI conference on artificial intelligence*.
- [11] Liu, K., Xu, L., Zhao, J. (2015). Co-extracting opinion targets and opinion words from online reviews based on the word alignment model. *IEEE Transactions on knowledge and data engineering*, 27 (3) 636-650.
- [12] Poria, S., Cambria, E., Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108, 42-49.
- [13] Shang, L., Wang, H., Dai, X., Zhang, M. (2012). Opinion target extraction for short comments. *PRICAI 2012: Trends in Artificial Intelligence*, p. 528-539.
- [14] Wang, W., Pan, S. J., Dahlmeier, D., Xiao, X. (2017). Coupled Multi-layer Attentions for Co-extraction of Aspect and Opinion Terms, *In: Proc of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*. 2017.
- [15] Yin, Y., Wei, F., Dong, L., Xu, K., Zhang, M., Zhou, M. (2016). Unsupervised word and dependency path embeddings for aspect term extraction. *In: Proc of 25th Inter. Joint Conf. on Artificial Intelligence (IJCAI'16)*, AAAI Press. 2979-2985.